Claims

What is claimed is:

Sub A1 1. A method for identifying groups of pages of common interest from a collection of hyper-linked pages, comprising the steps of:

identifying a plurality of community cores from the collection, each core being first and second sets of pages, each page in the first set pointing to every page in the second set; and

expanding each identified core into a full community, the full community being a subset of the pages regarding a particular topic.

2. The method as recited in claim 1, wherein:

the collection includes a plurality of sites, each site having one or more hyper-linked pages; and

the method further includes the step of removing the hyper-links between any two pages on a same site.

Sub A1 3. The method as recited in claim 2 further comprising the step of discarding the pages of predetermined sites.

4. The method as recited in claim 1 further comprising the steps of:

finding highly similar pages that have different names;

replacing the highly similar pages with a single representative page; and redirecting any hyper-links that pointed to one of the highly similar pages so that the redirected hyper-links now point to the representative page.

5. The method as recited in claim 1 further comprising the steps of:

discarding unnecessary pages from consideration to generate a set of candidate fan pages and a set of candidate center pages; and

using the set of candidate fan pages and set of candidate center pages as the first and second sets, respectively, to identify the community cores.

6. The method as recited in claim 5, wherein the step of discarding includes the steps of:

determining candidate fan pages, the candidate fan pages being those pointing to at least a predetermined number of different sites;

determining candidate center pages, the candidate center pages being those pointed to by one or more candidate fan pages; and

discarding all pages in the collection except the candidate fan pages and candidate center pages.

7. The method as recited in claim 6, wherein the determination of candidate fan pages is based on page content and the hyper-links pointing therefrom.

8. The method as recited in claim 5, wherein the step of identifying a plurality of community cores includes the step of finding a plurality of (i, j)-cores where i and j are the numbers of candidate fan pages and candidate center pages, respectively, that appear in each identified community core.

9. The method as recited in claim 8, wherein the step of finding a plurality of (i, j)-cores includes the steps of:

(a) discarding all candidate center pages that have fewer than i hyper-links pointing thereto;

(b) determining all candidate center pages that have i hyper-links pointing thereto and determining whether the i hyper-links represent a valid community core; and

(c) if the i hyper-links represent a valid community core, then outputting the valid core, otherwise, discarding the determined candidate center pages.

10. The method as recited in claim 9 further comprising the steps of:

(d) discarding all candidate fan pages that have fewer than j hyper-links pointing therefrom;

(e) determining all candidate fan pages that have j hyper-links pointing therefrom and determining whether the j hyper-links represent a valid community core; and

(f) if the j hyper-links represent a valid community core, then outputting the valid core, otherwise, discarding the determined candidate fan pages.

11. The method as recited in claim 10 further comprising the step of repeating steps (a)-(f) until every candidate fan page has more than j hyper-links pointing therefrom and every candidate center page has more than i hyper-links pointing thereto.

12. The method as recited in claim 10 further comprising the step of repeating steps (a)-(f) until a predetermined ending condition is satisfied.

13. The method as recited in claim 10 further comprising the steps of:

determining all (2,j) cores by examining all pairs of candidate fan pages;

for i = 3 to n, where n is a predetermined value:

(i) finding all (i,j)-cores by examining the (i-1,j)-cores; and

(ii) for each (i-1, j)-core, determining whether any of the candidate fan pages may be added to the (i-1, j)-core to yield a (i,j)-core; and

removing all (i,j)-cores that appear as subsets of (i',j) cores, where i' > i.

14. A computer program product for use with a computer system for identifying groups of pages of common interest from a collection of hyper-linked pages, the computer program product comprising:

a computer-readable medium;

means, provided on the computer-readable medium, for directing the system to identify a plurality of community cores from the collection, each core being first and second sets of pages, each page in the first set pointing to every page in the second set; and

means, provided on the computer-readable medium, for directing the system to expand each identified core into a full community, the full community being a subset of the pages regarding a particular topic.

15. The computer program product as recited in claim 14, wherein:

the collection includes a plurality of sites, each site having one or more hyper-linked pages; and

the product further includes means, provided on the computer-readable medium, for directing the system to remove the hyper-links between any two pages on a same site.

16. The computer program product as recited in claim 15 further comprising means, provided on the computer-readable medium, for directing the system to discard the pages of predetermined sites.

17. The computer program product as recited in claim 14 further comprising:

means, provided on the computer-readable medium, for directing the system to find highly similar pages that have different names;

means, provided on the computer-readable medium, for directing the system to replace the highly similar pages with a single representative page; and means, provided on the computer-readable medium, for directing the system to redirect any hyper-links that pointed to one of the highly similar pages so that the redirected hyper-links now point to the representative page.

18. The computer program product as recited in claim 14 further comprising:

means, provided on the computer-readable medium, for directing the system to discard unnecessary pages from consideration to generate a set of candidate fan pages and a set of candidate center pages; and

means, provided on the computer-readable medium, for directing the system to use the set of candidate fan pages and set of candidate center pages as the first and second sets, respectively, to identify the community cores.

19. The computer program product as recited in claim 18, wherein the means for directing to discard includes:

means, provided on the computer-readable medium, for directing the system to determine candidate fan pages, the candidate fan pages being those pointing to at least a predetermined number of different sites;

means, provided on the computer-readable medium, for directing the system to determine candidate center pages, the candidate center pages being those pointed to by one or more candidate fan pages; and means, provided on the computer-readable medium, for directing the system to discard all pages in the collection except the candidate fan pages and candidate center pages.

20. The computer program product as recited in claim 19, wherein the determination of candidate fan pages is based on page content and the hyper-links pointing therefrom.

21. The computer program product as recited in claim 18, the means for directing to identify a plurality of community cores includes means, provided on the computer-readable medium, for directing the system to find a plurality of (i, j)-cores where i and j are the numbers of candidate fan pages and candidate center pages, respectively, that appear in each identified community core.

22. The computer program product as recited in claim 21, wherein the means for directing to find a plurality of (i, j)-cores includes:

(a) means, provided on the computer-readable medium, for directing the system to discard all candidate center pages that have fewer than i hyper-links pointing thereto;

(b) means, provided on the computer-readable medium, for directing the system to determine all candidate center pages that have i hyper-links pointing thereto and determining whether the i hyper-links represent a valid community core; and

(c) means, provided on the computer-readable medium, for directing the system to output the valid core if the i hyper-links represent a valid community core, otherwise, to discard the determined candidate center pages.

23. The computer program product as recited in claim 22 further comprising:

(d) means, provided on the computer-readable medium, for directing the system to discard all candidate fan pages that have fewer than j hyper-links pointing therefrom;

(e) means, provided on the computer-readable medium, for directing the system to determine all candidate fan pages that have j hyper-links pointing therefrom and determining whether the j hyper-links represent a valid community core; and

(f) means, provided on the computer-readable medium, for directing the system to output the valid core if the j hyper-links represent a valid community core, otherwise, discard the determined candidate fan pages.

24. The computer program product as recited in claim 23, wherein the operation of means (a)-(f) is repeated until every candidate fan page has more than j hyper-links pointing therefrom and every candidate center page has more than i hyper-links pointing thereto.

25. The computer program product as recited in claim 23, wherein the operation of means (a)-(f) is repeated until a predetermined ending condition is satisfied.

26. The computer program product as recited in claim 23 further comprising:

means, provided on the computer-readable medium, for directing the system to determine all (2,j) cores by examining all pairs of candidate fan pages;

for i = 3 to n, where n is a predetermined value:

(i) means, provided on the computer-readable medium, for directing the system to find all (i,j)-cores by examining the (i-1,j)-cores; and

(ii) for each (i-1, j)-core, means, provided on the computer-readable medium, for directing the system to determine whether any of the candidate fan pages may be added to the (i-1, j)-core to yield a (i,j)-core; and

means, provided on the computer-readable medium, for directing the system to remove all (i,j)-cores that appear as subsets of (i',j) cores, where i' > i.

~~27.~~ ~~A system for identifying groups of pages of common interest from a collection of hyper-linked pages, comprising:~~

~~means for identifying a plurality of community cores from the collection, each core being first and second sets of pages, each page in the first set pointing to every page in the second set; and~~

~~means for expanding each identified core into a full community, the full community being a subset of the pages regarding a particular topic.~~

~~28. The system as recited in claim 27, wherein:~~

~~the collection includes a plurality of sites, each site having one or more hyper-linked pages; and~~

~~the method further includes the step of removing the hyper-links between any two pages on a same site.~~

29. The system as recited in claim 28 further comprising means for discarding the pages of predetermined sites.

30. The system as recited in claim 27 further comprising:

means for finding highly similar pages that have different names;

means for replacing the highly similar pages with a single representative page; and

means for redirecting any hyper-links that pointed to one of the highly similar pages so that the redirected hyper-links now point to the representative page.

31. The system as recited in claim 27 further comprising:

means for discarding unnecessary pages from consideration to generate a set of candidate fan pages and a set of candidate center pages; and

means for using the set of candidate fan pages and set of candidate center pages as the first and second sets, respectively, to identify the community cores.

32. The system as recited in claim 31, wherein the means for discarding includes:

means for determining candidate fan pages, the candidate fan pages being those pointing to at least a predetermined number of different sites;

means for determining candidate center pages, the candidate center pages being those pointed to by one or more candidate fan pages; and

means for discarding all pages in the collection except the candidate fan pages and candidate center pages.

33. The system as recited in claim 32, wherein the determination of candidate fan pages is based on page content and the hyper-links pointing therefrom.

34. The system as recited in claim 31, the means for identifying a plurality of community cores includes means for finding a plurality of (i, j)-cores where i and j are the numbers of candidate fan pages and candidate center pages, respectively, that appear in each identified community core.

35. The system as recited in claim 34, wherein the means for finding a plurality of (i, j)-cores includes:

(a) means for discarding all candidate center pages that have fewer than i hyper-links pointing thereto;

(b) means for determining all candidate center pages that have i hyper-links pointing thereto and determining whether the i hyper-links represent a valid community core; and

(c) means for outputting the valid core if the i hyper-links represent a valid community core, otherwise, discarding the determined candidate center pages.

36. The system as recited in claim 35 further comprising:

(d) means for discarding all candidate fan pages that have fewer than j hyper-links pointing therefrom;

(e) means for determining all candidate fan pages that have j hyper-links pointing therefrom and determining whether the j hyper-links represent a valid community core; and

(f) means for outputting the valid core if the j hyper-links represent a valid community core, otherwise, discarding the determined candidate fan pages.

37. The system as recited in claim 36, wherein the operation of means (a)-(f) is repeated until every candidate fan page has more than j hyper-links pointing therefrom and every candidate center page has more than i hyper-links pointing thereto.

38. The system as recited in claim 36, wherein the operation of means (a)-(f) is repeated until a predetermined ending condition is satisfied.

39. The system as recited in claim 36 further comprising:

means for determining all (2,j) cores by examining all pairs of candidate fan pages;

for i = 3 to n, where n is a predetermined value:

(i) means for finding all (i,j)-cores by examining the (i-1,j)-cores; and

(ii) for each (i-1, j)-core, means for determining whether any of the candidate fan pages may be added to the (i-1, j)-core to yield a (i,j)-core; and

means for removing all (i,j)-cores that appear as subsets of (i',j) cores, where i' > i.